The Causal Interpretations of Bayesian Hypergraphs

Zhiyu Wang[1]**, Mohammad Ali Javidian**[2]**, Linyuan Lu**[1]**, Marco Valtorta**[2]
[1]Department of Mathematics
[2]Department of Computer Science and Engineering
University of South Carolina, Columbia, SC, 29208

Abstract

We propose a causal interpretation of Bayesian hypergraphs (Javidian et al. 2018), a probabilistic graphical model whose structure is a directed acyclic hypergraph, that extends the causal interpretation of LWF chain graphs. We provide intervention formulas and a graphical criterion for intervention in Bayesian hypergraphs that specializes to a new graphical criterion for intervention in LWF chain graphs and sheds light on the causal interpretation of interaction as represented by undirected edges in LWF chain graphs or heads in Bayesian hypergraphs.

Introduction and Motivation

Probabilistic graphical models are graphs in which nodes represent random variables and edges represent conditional independence assumptions. They provide a compact way to represent the joint probability distributions of a set of random variables. In *undirected* graphical models, e.g., Markov networks (see (Darroch et al. 1980; Pearl 1988)), there is a simple rule for determining independence: two set of nodes A and B are conditionally independent given C if removing C separates A and B. On the other hand, *directed* graphical models, e.g. Bayesian networks (see (Kiiveri, Speed, and Carlin 1984; Wermuth and Lauritzen 1983; Pearl 1988)), which consist of a directed acyclic graph (DAG) and a corresponding set of conditional probability tables, have a more complicated rule (d-separation) for determining independence. More complex graphical models include various types of graphs with edges of several types (e.g., (Cox and Wermuth 1993; 1996; Richardson and Spirtes 2002; Peña 2014)), including chain graphs (Lauritzen and Wermuth 1989; Lauritzen 1996), for which different interpretations have emerged (Andersson, Madigan, and Perlman 1996; Drton 2009).

Probabilistic Graphical Models (PGMs) enjoy a well-deserved popularity because they allow explicit representation of structural constraints in the language of graphs and similar structures. From the perspective of efficient belief update, factorization of the joint probability distribution of random variables corresponding to variables in the graph is paramount, because it allows decomposition of the calculation of the evidence or of the posterior probability (Lauritzen and Jensen 1997). The proliferation of different PGMs that allow factorizations of different kinds leads us to consider a more general graphical structure in this paper, namely directed acyclic hypergraphs. Since there are many more hypergraphs than DAGs, undirected graphs, chain graphs, and, indeed, other graph-based networks, Bayesian hypergraphs can model much finer factorizations and thus are more computationally efficient. See (Javidian et al. 2018) for examples of how Bayesian hypergraphs can model graphically functional dependencies that are hidden in probability tables when using BNs or CGs. We provide a causal interpretation of Bayesian hypergraphs that extends the causal interpretation of LWF chain graphs (Lauritzen and Richardson 2002) by giving corresponding formulas and a graphical criterion for intervention, which operationalize the intuition that feedback processes in LWF chain graphs and Bayesian hypergraphs models turn into causal processes when variables in them are conditioned upon by intervention. The addition of feedback processes and its causal interpretation is a conceptual advance within the three-level causal hierarchy (Pearl 2009).

Bayesian Hypergraphs: Definitions

Hypergraphs are generalizations of graphs such that each edge is allowed to contain more than two vertices. Formally, an *(undirected) hypergraph* is a pair $\mathcal{H} = (V, \mathcal{E})$, where $V = \{v_1, v_2, \cdots, v_n\}$ is the set of *vertices* (or *nodes*) and $\mathcal{E} = \{h_1, h_2, \cdots, h_m\}$ is the set of *hyperedges* where $h_i \subseteq V$ for all $i \in [m]$. If

$|h_i| = k$ for every $i \in [m]$, then we say $\mathcal{H}$ is a *k-uniform* (undirected) hypergraph. A *directed hyperedge* or *hyperarc* h is an ordered pair, $h = (X, Y)$, of (possibly empty) subsets of V where $X \cap Y = \emptyset$; X is the called the *tail* of h while Y is the *head* of h. We write $X = T(h)$ and $Y = H(h)$. We say a directed hyperedge h is *fully directed* if none of $H(h)$ and $T(h)$ are empty. A *directed hypergraph* is a hypergraph such that all of the hyperedges are directed. A *(s,t)-uniform directed hypergraph* is a directed hypergraph such that the tail and head of every directed edge have size s and t respectively. For example, any DAG is a $(1,1)$-uniform hypergraph (but not vice versa). An undirected graph is a $(0,2)$-uniform hypergraph. Given a hypergraph $\mathcal{H}$, we use $V(\mathcal{H})$ and $E(\mathcal{H})$ to denote the the vertex set and edge set of $\mathcal{H}$ respectively.

We say two vertices u and v are *co-head* (or *co-tail*) if there is a directed hyperedge h such that $\{u, v\} \subset H(h)$ (or $\{u, v\} \subset T(h)$ respectively). Given another vertex $u \neq v$, we say u is a *parent* of v, denoted by $u \rightarrow v$, if there is a directed hyperedge h such that $u \in T(h)$ and $v \in H(h)$. If u and v are co-head, then u is a *neighbor* of v. If u, v are neighbors, we denote them by $u - v$. Given $v \in V$, we define parent ($pa(v)$), neighbor ($nb(v)$), boundary ($bd(v)$), ancestor ($an(v)$), anterior ($ant(v)$), descendant ($de(v)$), and non-descendant ($nd(v)$) for hypergraphs exactly the same as for graphs (and therefore use the same names). The same holds for the equivalent concepts for $\tau \subseteq V$. Note that it is possible that some vertex u is both the parent and neighbor of v.

A *partially directed cycle* in $\mathcal{H}$ is a sequence $\{v_1, v_2, \ldots v_k\}$ satisfying that v_i is either a neighbor or a parent of v_{i+1} for all $1 \leq i \leq k$ and $v_i \rightarrow v_{i+1}$ for some $1 \leq i \leq k$. Here $v_{k+1} \equiv v_1$. We say a directed hypergraph $\mathcal{H}$ is *acyclic* if $\mathcal{H}$ contains no partially directed cycle. For ease of reference, we call a directed acyclic hypergraph a *DAH* or a *Bayesian hypergraph structure* (as defined in Definition 1). Note that for any two vertices u, v in a directed acyclic hypergraph $\mathcal{H}$, u can not be both the parent and neighbor of v otherwise we would have a partially directed cycle.

Remark 1. *DAHs are generalizations of undirected graphs, DAGs and chain graphs. In particular an undirected graph can be viewed as a DAH in which every hyperedge is of the form $(\emptyset, \{u, v\})$. A DAG is a DAH in which every hyperedge is of the form $(\{u\}, \{v\})$. A chain graph is a DAH in which every hyperedge is of the form $(\emptyset, \{u, v\})$ or $(\{u\}, \{v\})$.*

We define the *chain components* of $\mathcal{H}$ as the equivalence classes under the equivalence relation where two vertices v_1, v_t are equivalent if there exists a sequence of distinct vertices $v_1, v_2, \ldots, v_t$ such that v_i and v_{i+1} are co-head for all $i \in [t-1]$. The chain components $\{\tau : \tau \in \mathcal{D}\}$ yields an unique natural partition of the vertex set $V(\mathcal{H}) = \bigcup_{\tau \in \mathcal{D}} \tau$ with the following properties:

Proposition 1. *Let $\mathcal{H}$ be a DAH and $\{\tau : \tau \in \mathcal{D}\}$ be its chain components. Let G be a graph obtained from $\mathcal{H}$ by contracting each element of $\{\tau : \tau \in \mathcal{D}\}$ into a single vertex and creating a directed edge from $\tau_i \in V(G)$ to $\tau_j \in V(G)$ in G if and only if there exists a hyperedge $h \in E(\mathcal{H})$ such that $T(h) \cap \tau_i \neq \emptyset$ and $H(h) \cap \tau_j \neq \emptyset$. Then G is a DAG.*

Proof. See (Javidian et al. 2018). □

Note that the DAG obtained in Proposition 1 is unique and given a DAH $\mathcal{H}$ we call such G the *canonical DAG* of $\mathcal{H}$.

Definition 1. *A Bayesian hypergraph is a triple $(V, \mathcal{H}, P)$ such that V is a set of random variables, $\mathcal{H}$ is a DAH on the vertex set V and P is a multivariate probability distribution on V such that the local Markov property holds with respect to the DAH $\mathcal{H}$, i.e., for any vertex $v \in V(\mathcal{H})$,*

$$v \perp\!\!\!\perp nd(v) \backslash cl(v) \mid bd(v). \tag{1}$$

For a Bayesian hypergraph $\mathcal{H}$ whose underlying DAH is a LWF DAH, we call $\mathcal{H}$ a *LWF Bayesian hypergraph*.

Bayesian hypergraphs factorizations

The factorization of a probability measure P according to a Bayesian hypergraph is similar to that of a chain graph. Before we present the factorization property, let us introduce some additional terminology. Given a DAH $\mathcal{H}$, we use $\mathcal{H}^u$ to denote the undirected hypergraph obtained from $\mathcal{H}$ by replacing each directed hyperedge $h = (A, B)$ of $\mathcal{H}$ into an undirected hyperedge $A \cup B$. Given a family of sets $\mathcal{F}$, define a partial order $(\mathcal{F}, \leq)$ on $\mathcal{F}$ such that for two sets $A, B \in \mathcal{F}$, $A \leq B$ if and only if $A \subseteq B$. Let $\mathcal{M}(\mathcal{F})$ denote the set of maximal elements in $\mathcal{F}$, i.e., no element in $\mathcal{M}(\mathcal{F})$ contains another element as subset. When $\mathcal{F}$ is a set of directed hyperedges, we abuse the notation to denote $\mathcal{M}(\mathcal{F}) = \mathcal{M}(\mathcal{F}^u)$. Let $\mathcal{H}$ be a directed acyclic hypergraph and $\{\tau : \tau \in \mathcal{D}\}$ be its chain components. Assume that a probability distribution P has a density f, with respect to some product measure $\mu = \times_{\alpha \in V} \mu_\alpha$ on $\mathcal{X} = \times_{\alpha \in V} \mathcal{X}_\alpha$. Now we say a probability measure P *factorizes* according to $\mathcal{H}$ if it has density f such that

(i) f factorizes as in the directed acyclic case:

$$f(x) = \prod_{\tau \in \mathcal{D}} f(x_\tau \mid x_{pa(\tau)}). \tag{2}$$

(ii) For each $\tau \in \mathcal{D}$, define $\mathcal{H}^*_\tau$ to be the subhypergraph of $\mathcal{H}_{\tau \cup pa(\tau)}$ containing all edges h in $\mathcal{H}_{\tau \cup pa(\tau)}$ such that $H(h) \subseteq \tau$.

$$f(x_\tau \mid x_{pa(\tau)}) = \prod_{h\in\mathcal{M}(\mathcal{H}^*_\tau)} \psi_h(x). \qquad (3)$$

where ψ_h are non-negative functions depending only on x_h and $\int_{X_\tau} \prod_{h\in\mathcal{M}(\mathcal{H}^*_\tau)} \psi_h(x)\mu_\tau(dx_\tau) = 1$. Equivalently, we can also write $f(x_\tau \mid x_{pa(\tau)})$ as

$$f(x_\tau \mid x_{pa(\tau)}) = Z^{-1}(x_{pa(\tau)}) \prod_{h\in\mathcal{M}(\mathcal{H}^*_\tau)} \psi_h(x), \qquad (4)$$

where $Z^{-1}(x_{pa(\tau)}) = \int_{X_\tau} \prod_{h\in\mathcal{M}(\mathcal{H}^*_\tau)} \psi_h(x)\mu_\tau(dx_\tau)$.

Remark 2. *One of the key advantages of Bayesian hypergraphs is that they allow much finer factorizations of probability distributions compared to chain graph models. We will illustrate with a simple example in Figure 1.*

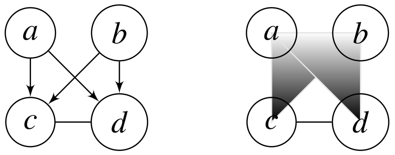

Figure 1: (1) a chain graph G; (2) a Bayesian hypergraph $\mathcal{H}$.

Note that in Figure 1 (1), the factorization according to G is

$$\begin{aligned} f(x) &= f(x_a)f(x_b)f(x_{cd} \mid x_{ab}) \\ &= f(x_a)f(x_b)\psi_{abcd}(x) \end{aligned}$$

In Figure 1 (2), the factorization according to $\mathcal{H}$ is

$$\begin{aligned} f(x) &= f(x_a)f(x_b)f(x_{cd} \mid x_{ab}) \\ &= f(x_a)f(x_b)\psi_{abc}(x)\psi_{abd}(x)\psi_{cd}(x) \end{aligned}$$

Note that although G and $\mathcal{H}$ have the same global Markov properties, the factorization according to $\mathcal{H}$ is one step further compared to the factorization according to G. Suppose each of the variables of $\{a, b, c, d\}$ can take k values. Then the factorization according to G will require a conditional probability table of size k^4 while the factorization according to $\mathcal{H}$ only needs a table of size $\Theta(k^3)$ asymptotically. Hence, a Bayesian hypergraph model allows much finer factorizations and thus achieves higher memory efficiency.

Remark 3. *We remark that the factorization formula defined in* (3) *is in fact the most general possible in the sense that it allows all possible factorizations of a probability distribution admitted by a DAH. In particular, given a Bayesian hypergraph $\mathcal{H}$ and one of its chain components τ, the factorization scheme in* (3) *allows a distinct function for each maximal subset of $\tau \cup pa_{\mathfrak{D}}(\tau)$ that intersects τ ($pa_{\mathfrak{D}}$ is the parent of τ in the canonical DAG of $\mathcal{H}$). For each subset S of $\tau \cup pa_{\mathfrak{D}}(\tau)$ that does not intersect τ, recall that the factorization in* (3) *can be rewritten as follows:*

$$f(x_\tau \mid x_{pa(\tau)}) = \frac{\prod_{h\in\mathcal{M}(\mathcal{H}^*_\tau)} \psi_h(x)}{\int_{X_\tau} \prod_{h\in\mathcal{M}(\mathcal{H}^*_\tau)} \psi_h(x)\mu_\tau(dx_\tau)}.$$

Observe that $\psi_S(x)$ is a function that does not depend on values of variables in τ. Hence $\psi_S(x)$ can be factored out from the integral above and cancels out with itself in $f(x_\tau \mid x_{pa(\tau)})$. Thus, the factorization formula in (3) *or* (4) *in fact allows distinct functions for all possible maximal subsets of $\tau \cup pa_{\mathfrak{D}}(\tau)$.*

Intervention in Bayesian hypergraphs

Formally, intervention in Bayesian hypergraphs can be defined analogously to intervention in LWF chain graphs (Lauritzen and Richardson 2002). In this section, we give graphical procedures that are consistent with the intervention formulas for chain graphs (Equation (5), (6)) and for Bayesian hypergraphs (Equation (7), (8)). Before we present the details, we need some additional definitions and tools to determine when factorizations according to two chain graphs or DAHs are equivalent in the sense that they could be written as products of the same type of functions (functions that depend on same set of variables). We say two chain graphs G_1, G_2 admit the same *factorization decomposition* if for every probability density f that factorizes according to G_1, f also factorizes according to G_2, and vice versa. Similarly, two DAHs $\mathcal{H}_1, \mathcal{H}_2$ admit the same *factorization decomposition* if for every probability density f that factorizes according to $\mathcal{H}_1$, f also factorizes according to $\mathcal{H}_2$, and vice versa.

Factorization equivalence and intervention in LWF chain graphs

In this subsection, we will give graphical procedures to model intervention based on the formula introduced in (Lauritzen and Richardson 2002). Let us first give some background. In many statistical context, we would like to modify the distribution of a variable Y by intervening externally and forcing the value of another variable X to be x. This is commonly referred as *conditioning by intervention* or *conditioning by action* and denoted by $Pr(y \| x)$ or $Pr(y \mid X \leftarrow x)$. Other expressions such as $Pr(Y_x = y), P_{man(x)}(y), set(X = x), X = \hat{x}$ or $do(X = x)$ have also been used to denote intervention conditioning

(Splawa-Neyman 1990; Rubin 1974; Pearl 1993; 1995; 2009).

Let G be a chain graph with chain components $\{\tau : \tau \in \mathcal{D}\}$. Moreover, assume further that a subset A of variables in $V(G)$ are set such that for every $a \in A$, $x_a = a_0$. Lauritzen and Richardson, in (Lauritzen and Richardson 2002), generalized the conditioning by intervention formula for DAGs and gave the following formula for intervention in chain graphs (where it is understood that the probability of any configuration of variables inconsistent with the intervention is zero). A probability density f factorizes according to G (with A intervened) if

$$f(x \| x_A) = \prod_{\tau \in \mathcal{D}} f(x_{\tau \backslash A} \mid x_{pa(\tau)}, x_{\tau \cap A}). \quad (5)$$

Moreover, for each $\tau \in \mathcal{D}$,

$$f(x_{\tau \backslash A} \,|\, x_{pa(\tau)}, x_{\tau \cap A}) = Z^{-1}(x_{pa(\tau)}, x_{\tau \cap A}) \prod_{h \in C} \psi_h(x) \quad (6)$$

where C is the set of maximal cliques in $(G_{\tau \cup pa(\tau)})^m$ and $Z^{-1}(x_{pa(\tau)}, x_{\tau \cap A}) = \int_{X_{\tau \backslash A}} \prod_{h \in C} \psi_h(x) \mu_{\tau \backslash A}(dx_{\tau \backslash A})$.

Definition 2. *G_1 and G_2 be two chain graphs. Given a subset $A_1 \subseteq V(G_1)$ and $A_2 \subseteq V(G_2)$, we say (G_1, A_1) and (G_2, A_2) are* factorization-equivalent[1] *if they become the same chain graph after removing from G_i all vertices in A_i together with the edges incident to vertices in A_i for $i \in \{1, 2\}$.*

Typically, A_i in Definition 2 is a set of constant variables in $V(G_i)$ created by intervention.

Theorem 1. *Let G_1 and G_2 be two chain graphs defined on the same set of variables V. Moreover a common set of variables A in V are set by intervention such that for every $a \in A$, $x_a = a_0$. If (G_1, A) and (G_2, A) are factorization-equivalent, then G_1 and G_2 admit the same factorization decomposition.*

Proof. Let G_0 be the chain graph obtained from G_1 by removing all vertices in A and the edges incident to A. It suffices to show that G_1 and G_2 both admit the same factorization decomposition as G_0. Let $\mathcal{D}_1, \mathcal{D}_0$ be the set of chain components of G_1 and G_0 respectively. Let $\tau \in \mathcal{D}_1$ be an arbitrary chain component of G_1. By the factorization formula in (6), it follows that

$$f(x_{\tau \backslash A} \mid x_{pa(\tau)}, x_{\tau \cap A}) = Z^{-1}(x_{pa(\tau)}, x_{\tau \cap A}) \prod_{h \in C} \psi_h(x)$$

where C is the set of maximal cliques in $(G_{\tau \cup pa(\tau)})^m$ and $Z^{-1}(x_{pa(\tau)}, x_{\tau \cap A}) = \int_{X_{\tau \backslash A}} \prod_{h \in C} \psi_h(x) \mu_{\tau \backslash A}(dx_{\tau \backslash A})$.

Notice that for any maximal clique $h_1 \in C$ such that $h_1 \cap A = \emptyset$, h_1 is also a clique in $(G_0[\tau \backslash A])^m$. For $h_1 \in C$ with $h_1 \cap A \neq \emptyset$, there are two cases:

Case 1: $(h_1 \cap \tau) \backslash A \neq \emptyset$. In this case, observe that $h_1 \backslash A$ is also a clique in $(G_0[\tau \backslash A])^m$, thus is contained in some maximal clique h' in $(G_0[\tau \backslash A])^m$. Since all variables in A are pre-set as constants, it follows that $\psi_{h_1}(x)$ also appears in a factor in the factorization of f according to G_0.

Case 2: $h_1 \cap \tau \subseteq A$. In this case, note that $h_1 \cap \tau$ is disjoint with $\tau \backslash A$. Hence $\psi_{h_1}(x)$ appears as a factor independently of $x_{\tau \backslash A}$ in both $Z^{-1}(x_{pa(\tau)}, x_{\tau \cap A})$ and $\prod_{h \in C} \psi_h(x)$, which cancels out with itself.

Thus it follows that every probability density f that factorizes according to G_1 also factorizes according to G_0. On the other hand, it is easy to see that for every $\tau' \in \mathcal{D}_0$ and every maximal clique h' in $(G_0[\tau'])^m$, h' is contained in some maximal clique h in $(G_1[\tau])^m$ for some $\tau \in \mathcal{D}_1$. Hence we can conclude that G_1 and G_0 admit the same factorization decomposition. The above argument also works for G_2 and G_0. Thus, G_1 and G_2 admit the same factorization decomposition. □

We now define a graphical procedure (call it *redirection procedure*) that is consistent with the intervention formula in Equation (5) and (6). Let G be a chain graph. Given an intervened set of variables $A \subseteq V(G)$, let $\hat{G}$ be the chain graph obtained from G by performing the following operation: for every $u \in A$ and every undirected edge $e = \{u, w\}$ containing u, replace e by a directed edge from u to w; finally remove all the directed edges that point to some vertex in A. By replacing the undirected edge with a directed edge, we replace any feedback mechanisms that include a variable in A with a causal mechanism. The intuition behind the procedure is the following. Since a variable that is set by intervention cannot be modified, the symmetric feedback relation is turned into an asymmetric causal one. Similarly, we can justify this graphical procedure as equivalent to "striking out" some equations in the Gibbs process on top of p. 338 of (Lauritzen and Richardson 2002), as Lauritzen and Richardson (Richardson 2018) did for Equation (18) in (Lauritzen and Richardson 2002).

Theorem 2. *Let G be a chain graph with a subset of variables $A \subseteq V(G)$ set by intervention such that for every $a \in A$. $x_a = a_0$. Let $\hat{G}$ be obtained from G by the redirection procedure. Then G and $\hat{G}$ admit the same factorization decomposition.*

Proof. It is not hard to see that removing from $\hat{G}$ and G all vertices in A and all edges incident to A results

[1] This term was defined for a different purpose in (Studený 1992).

in the same chain graph. Hence by Theorem (1), G and $\hat{G}$ admit the same factorization decomposition. □

Example 1. *Consider the chain graph G shown in Figure 2. Let $\hat{G}$ be the graph obtained from G through the redirection procedure described in this subsection. Let G_0 be the chain graph obtained from G by deleting the vertex c_0 and the edges incident to c_0. We will compare the factorization decomposition according to the formula* (5),(6) *as well as the graph structure $\hat{G}$ and G_0.*

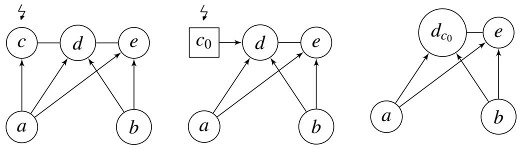

Figure 2: (a) A chain graph G; (b) The graph $\hat{G}$ obtained from G through the redirection procedure; (c) The graph G_0 obtained from G by deleting variables in A.

By formulas (5) *and* (6) *proposed in (Lauritzen and Richardson 2002), when x_c is set as c_0 by intervention,*

$$\begin{aligned} f(x \| x_c) &= f(x_a) f(x_b) f(x_{de} \mid x_{abc_0}) \\ &= f(x_a) f(x_b) \frac{\psi_{ac_0d}(x)\psi_{abde}(x)}{\sum_{d,e} \psi_{ac_0d}(x)\psi_{abde}(x)}. \end{aligned}$$

Now consider the factorization according to $\hat{G}$. The chain components of $\hat{G}$ are $\{\{a\},\{b\},\{c\},\{d,e\}\}$ with x_c set to be c_0. The factorization according to $\hat{G}$ is as follows:

$$\begin{aligned} f_{\hat{G}}(x \| x_c) &= f_{\hat{G}}(x_a) f_{\hat{G}}(x_b) f_{\hat{G}}(x_c) f_{\hat{G}}(x_{de} \mid x_{abc_0}) \\ &= f_{\hat{G}}(x_a) f_{\hat{G}}(x_b) f_{\hat{G}}(x_c) \frac{\psi_{ac_0d}(x)\psi_{abde}(x)}{\sum_{d,e} \psi_{ac_0d}(x)\psi_{abde}(x)}, \end{aligned}$$

where $f(x_c) = 1$ when $x_c = c_0$ and otherwise 0. *Hence G and $\hat{G}$ admit the same factorization.*

Now consider the factorization according to G_0. The chain components of G_0 are $\{\{a\},\{b\},\{d,e\}\}$. The factorization according to G_0 is as follows:

$$\begin{aligned} f_0(x) &= f_0(x_a) f_0(x_b) f_0(x_{de} \mid x_{ab}) \\ &= f_0(x_a) f_0(x_b) \frac{\psi_{ad}(x)\psi_{abde}(x)}{\sum_{d,e} \psi_{ad}(x)\psi_{abde}(x)}, \end{aligned}$$

Observe that $f_0(x)$ has the same form of decomposition as $f(x \| x_c)$ since x_c is set to be c_0 in $\psi_{ac_0d}(x)$ (with the understanding that the probability of any configuration of variables with $x_c \neq c_0$ is zero). Hence we can conclude that $G, \hat{G}$ (with x_c intervened) and G_0 admit the same factorization decomposition.

Factorization equivalence and intervention in Bayesian hypergraphs

Intervention in Bayesian hypergraphs can be modeled analogously to the case of chain graphs. We use the same notation as before. Let $\mathcal{H}$ be a DAH and $\{\tau : \tau \in \mathcal{D}\}$ be its chain components. Moreover, assume further that a subset A of variables in $V(\mathcal{H})$ are set such that for every $a \in A$, $x_a = a_0$. Then a probability density f factorizes according to $\mathcal{H}$ (with A intervened) as follows: (where it is understood that the probability of any configuration of variables inconsistent with the intervention is zero):

$$f(x \| x_A) = \prod_{\tau \in \mathcal{D}} f(x_{\tau \setminus A} \mid x_{pa(\tau)}, x_{\tau \cap A}). \tag{7}$$

For each $\tau \in \mathcal{D}$, define $\mathcal{H}^*_\tau$ to be the subhypergraph of $\mathcal{H}_{\tau \cup pa_{\mathcal{D}}(\tau)}$ containing all edges h in $\mathcal{H}_{\tau \cup pa(\tau)}$ such that $H(h) \subseteq \tau$, then

$$f(x_{\tau \setminus A} \,|\, x_{pa(\tau)}, x_{\tau \cap A}) = Z^{-1}(x_{pa(\tau)}, x_{\tau \cap A}) \prod_{h \in \mathcal{M}(\mathcal{H}^*_\tau)} \psi_h(x) \tag{8}$$

where

$$Z^{-1}(x_{pa(\tau)}, x_{\tau \cap A}) = \int_{\mathcal{X}_{\tau \setminus A}} \prod_{h \in \mathcal{M}(\mathcal{H}^*_\tau)} \psi_h(x) \mu_{\tau \setminus A}(dx_{\tau \setminus A})$$

and ψ_h are non-negative functions that depend only on x_h.

Definition 3. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Bayesian hypergraphs. Given a subset of variables $A_1 \subseteq V(\mathcal{H}_1)$ and $A_2 \subseteq V(\mathcal{H}_2)$, we say $(\mathcal{H}_1, A_1)$ and $(\mathcal{H}_2, A_2)$ are* factorization-equivalent *if performing the following operations to $\mathcal{H}_1$ and $\mathcal{H}_2$ results in the same directed acyclic hypergraph:*

(i) Deleting all hyperedges with empty head, i.e., hyperedges of the form $(S, \emptyset)$.

(ii) Deleting every hyperedge that is contained in some other hyperedge, i.e., delete h if there is another h' such that $T(h) \subseteq T(h')$ and $H(h) \subseteq H(h')$.

(iii) Shrinking all hyperedges of $\mathcal{H}_i$ containing vertices in A_i, i.e. replace every hyperedge h of $\mathcal{H}_i$ by $h' = (T(h) \setminus A_i, H(h) \setminus A_i)$ for $i \in \{1,2\}$.

Typically, A_i in Definition 3 is a set of constant variables in V created by intervention.

Theorem 3. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two DAHs defined on the same set of variables V. Moreover, a common set of variables A in V are set by intervention such that for every $a \in A$, $X_a = a_0$. If $(\mathcal{H}_1, A)$ and $(\mathcal{H}_2, A)$ are factorization-equivalent, then $\mathcal{H}_1$ and $\mathcal{H}_2$ admit the same factorization decomposition.*

Proof. Similar to the proof in Theorem 1, let $\mathcal{H}_0$ be the DAH obtained from $\mathcal{H}_1$ (or $\mathcal{H}_2$) by performing the operations above repeatedly. Let $\mathcal{D}_1$ and $\mathcal{D}_0$ be the set of chain components of $\mathcal{H}_1$ and $\mathcal{H}_0$ respectively. First, note that performing the operation (i) does not affect the factorization since hyperedges of the form $h = (S, \emptyset)$ never appear in the factorization decomposition due to the fact that $H(h) \cap \tau = \emptyset$ for every $\tau \in \mathcal{D}_1$. Secondly, (ii) does not change the factorization decomposition too since if one hyperedge h is contained in another hyperedge h' as defined, then $\psi_h(x)$ can be simply absorbed into $\psi_{h'}(x)$ by replacing $\psi_{h'}(x)$ with $\psi_{h'}(x) \cdot \psi_h(x)$.

Now let $\tau \in \mathcal{D}_1$ be an arbitrary chain component of $\mathcal{H}_1$ and $h_1 \in \mathcal{H}_1[\tau]^*$, i.e., the set of hyperedges in $\mathcal{H}_1$ whose head intersects τ. Suppose that τ is separated into several chain components $\tau'_1, \tau'_2, \cdots, \tau'_t$ in $\mathcal{H}_0$ because of the shrinking operation. If $h_1 \cap A = \emptyset$, then h_1 is also a hyperedge in $\mathcal{H}_0[\tau \backslash A]^*$. If $h_1 \cap A \neq \emptyset$, there are two cases:

Case 1: $H(h_1) \subseteq A$. Then since variables in A are constants, it follows that in Equation (8), $\psi_{h_1}(x)$ does not depend on variables in $\tau \backslash A$. Hence $\psi_h(x)$ appears as factors independent of variables in $\tau \backslash A$ in both $Z^{-1}(x_{pa(\tau)}, x_{\tau\cap A})$ and $\prod_{h\in\mathcal{M}(\mathcal{H}_\tau^*)} \psi_h(x)$, thus cancels out with itself. Note that, h_1 does not exist in $\mathcal{H}_0$ too since h_1 becomes a hyperedge with empty head after being shrinked and thus is deleted in Operation (i).

Case 2: $H(h_1) \backslash A \neq \emptyset$. In this case, $H(h_1) \backslash A$ must be entirely contained in one of $\{\tau'_1, \cdots, \tau'_t\}$. Without loss of generality, say $H(h_1) \backslash A \subseteq \tau'_1$ in $\mathcal{H}_0$. Then note that $h_1 \backslash A$ must be contained in some maximal hyperedge h' in $E(\mathcal{H}_0)$ such that $H(h') \cap \tau'_1 \neq \emptyset$. Moreover, recall that variables in A are constants. Hence ψ_{h_1} must appear in some factor in the factorization of f according to $\mathcal{H}_0$.

Thus it follows that every probability density f that factorizes according to $\mathcal{H}_1$ also factorizes according to $\mathcal{H}_0$. On the other hand, it is not hard to see that for every $\tau' \in \mathcal{D}_0$ and every hyperedge h' in $(\mathcal{H}_0[\tau'])^*$, h' is contained in some maximal hyperedge h in $(\mathcal{H}_1[\tau])^*$ for some $\tau \in \mathcal{D}_1$. Hence we can conclude that $\mathcal{H}_1$ and $\mathcal{H}_0$ admit the same factorization decomposition. The above argument also works for $\mathcal{H}_2$ and $\mathcal{H}_0$. Thus, $\mathcal{H}_1$ and $\mathcal{H}_2$ admit the same factorization decomposition. □

We now present a graphical procedure (call it *redirection procedure*) for modeling intervention in Bayesian hypergraph. Let $\mathcal{H}$ be a DAH and $\{\tau : \tau \in \mathcal{D}\}$ be its chain components. Suppose a set of variables x_A is set by intervention. We then modify $\mathcal{H}$ as follows: for each hyperedge $h \in E(\mathcal{H})$ such as $S = H(h) \cap A \neq \emptyset$, replace the hyperedge h by $h' = (T(h) \cup S, H(h) \backslash S)$. If a hyperedge has empty set as its head, delete that hyperedge. Call the resulting hypergraph $\hat{\mathcal{H}}_A$. We will show that the factorization according to $\hat{\mathcal{H}}_A$ is consistent with Equation (8).

Theorem 4. *Let $\mathcal{H}$ be a Bayesian hypergraph and $\{\tau : \tau \in \mathcal{D}\}$ be its chain components. Given an intervened set of variables x_A, let $\hat{\mathcal{H}}_A$ be the DAH obtained from $\mathcal{H}$ by replacing each hyperedge $h \in E(\mathcal{H})$ satisfying $S = H(h) \cap A \neq \emptyset$ by the hyperedge $h' = (T(h) \cup S, H(h) \backslash S)$ and removing hyperedges with empty head. Then $\mathcal{H}$ and $\hat{\mathcal{H}}$ admit the same factorization decomposition.*

Proof. This is a corollary of Theorem (3) since performing the operations (i)(ii)(iii) in the definition of factorization-equivalence of DAH to $\mathcal{H}$ and $\hat{\mathcal{H}}$ results in the same DAH. □

Example 2. *Let G be a chain graph as shown in Figure 3(a) and $\mathcal{H}$ be the canonical LWF Bayesian hypergraph of G as shown in Figure 3(b), constructed based on the procedure in (Javidian et al. 2018, Section 2.4). $\mathcal{H}$ has two directed hyperedges $(\{a\}, \{c, d\})$ and $(\{a, b\}, \{d, e\})$. Applying the redirection procedure for intervention in Bayesian hypergraphs leads to the Bayesian hypergraph $\hat{\mathcal{H}}$ in Figure 3(c). We show that using equations* (5) *and* (6) *for Figure 3(a) leads to the same result as if one uses the factorization formula for the Bayesian hypergraph in Figure 3(c). First, we compute $f(x||x_c)$ for chain graph in Figure 3(a). Based on equation* (5) *we have:*

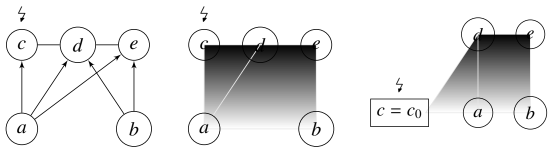

Figure 3: (a) A chain graph G; (b) the canonical LWF DAH $\mathcal{H}$ of G; (c) the resulting hypergraph $\hat{\mathcal{H}}$ after performing the graphical procedure on $\mathcal{H}$ when the variable c is intervened.

$$f(x||x_c) = f(x_a) f(x_b) f(x_{de} \mid x_{abc_0}),$$

as the effect of the atomic intervention $do(X_c = c_0)$. Then, using equation (6) *gives:*

$$f(x||x_c) = f(x_a) f(x_b) \frac{\psi_{ac_0d}(x)\psi_{abde}(x)}{\sum_{d,e} \psi_{ac_0d}(x)\psi_{abde}(x)}. \quad (9)$$

Now, we compute $f(x)$ for Bayesian hypergraph in Figure 3(c). Using equation (2) gives:

$$f(x||x_c) = f(x_a) f(x_b) f(x_{de} \mid x_{abc_0}).$$

Applying formula (3) gives:

$$f(x\|x_c) = f(x_a)f(x_b)f(x_c)\frac{\psi_{ac_0d}(x)\psi_{abde}(x)}{\sum_{d,e}\psi_{ac_0d}(x)\psi_{abde}(x)}. \quad (10)$$

Note that $f(x_c) = 1$, when $x_c = c_0$, otherwise $f(x_c) = 0$. As a result, the right side of equations (9) and (10) are the same.

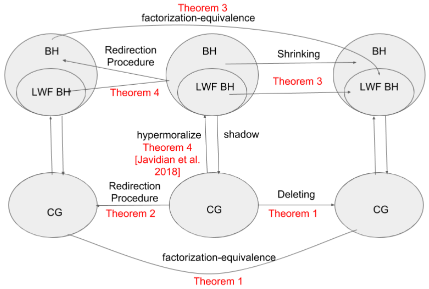

Figure 4: Commutative diagram of factorization equivalence

Remark 4. *Figure 4 summarizes our results. Given a chain graph G and its canonical LWF DAH $\mathcal{H}$, Theorem 4 in (Javidian et al. 2018) shows that G and $\mathcal{H}$ admit the same factorization decomposition. Suppose a set of variables A is set by intervention. Theorem 1 and 2 show that the the DAH obtained from G by the redirection procedure or deleting the variables in A admit the same factorization decomposition, which is also consistent with the intervention formula introduced in (Lauritzen and Richardson 2002). Similarly, Theorem 3 and 4 show that the DAH obtained from $\mathcal{H}$ by the redirection procedure or shrinking the variables in A admit the same factorization decomposition, which is consistent with a hypergraph analogue of the formula in (Lauritzen and Richardson 2002).*

Acknowledgments. This work is primarily supported by Office on Naval Research grant ONR N00014-17-1-2842.

References

Andersson, S. A.; Madigan, D.; and Perlman, M. D. 1996. Alternative markov properties for chain graphs. *Uncertainty in artificial intelligence* 40–48.

Cox, D. R., and Wermuth, N. 1993. Linear dependencies represented by chain graphs. *Statistical Science* 8(3):204–218.

Cox, D. R., and Wermuth, N. 1996. *Multivariate Dependencies-Models, Analysis and Interpretation*. Chapman and Hall.

Darroch, J. N.; Lauritzen, S. L.; ; and Speed, T. P. 1980. Markov fields and log-linear interaction models for contingency tables. *Scandinavian Journal of Statistics* 8(3):522–539.

Drton, M. 2009. Discrete chain graph models. *Bernoulli* 15(3):736–753.

Javidian, M. A.; Lu, L.; Valtorta, M.; and Wang, Z. 2018. On a hypergraph probabilistic graphical model. `https://arxiv.org/abs/1811.08372`.

Kiiveri, H.; Speed, T. P.; and Carlin, J. B. 1984. Recursive causal models. *Journal of the Australian Mathematical Society* 36:30–52.

Lauritzen, S., and Jensen, F. V. 1997. Local computations with valuations from a commutative semigroup. *Annals of Mathematics and Artificial Intelligence* 21:51–69.

Lauritzen, S., and Richardson, T. 2002. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 64(3):321–348.

Lauritzen, S., and Wermuth, N. 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics* 17(1):31–57.

Lauritzen, S. 1996. *Graphical Models*. Oxford Science Publications.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

Pearl, J. 1993. [bayesian analysis in expert systems]: Comment: Graphical models, causality and intervention. *Statistical Science* 8(3):266–269.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika* 82(4):669–710.

Pearl, J. 2009. *Causality. Models, reasoning, and inference*. Cambridge University Press.

Peña, J. M. 2014. Marginal amp chain graphs. *International Journal of Approximate Reasoning* 55:1185–1206.

Richardson, T. S., and Spirtes, P. 2002. Ancestral graph markov models. *The Annals of Statistics* 30(4):962–1030.

Richardson, T. S. 2018. personal communication.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688–701.

Splawa-Neyman, J. 1990. On the application of probability theory to agricultural experiments: Essay on principles. *Statistical Science* 5(4):465–472.

Studený, M. 1992. Conditional independence relations have no finite complete characterization. In Kubík, S., and Víšek, J., eds., *Information Theory, Statistical Decision Functions and Random Processes: Proceedings of the 11th Prague Conference - B*, volume 15, 377–396.

Wermuth, N., and Lauritzen, S. L. 1983. Graphical and recursive models for contingency tables. *Biometrika* 70(3):537–552.